VIT

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Summarizing Online Customer Reviews using Topic Modelling

Review - III

Team Number - 17

Team members:

SANJAY VENUGOPAL	– 20BCE2710
RAJVEER HEERA	– 20BCE0921
ZAMAN SALEEL	– 20BCE2025

Under the Guidance of:-

Prof Rajeshkannan R

1. Abstract

We frequently come across huge amounts of texts that need to be processed to obtain useful information, but before reading the documents, it would be a lot simpler if we could acquire a general idea of what they are about. We can shortlist and categorize the comments based on topics of interest without having to read through all of them if we have a lot of documents and can identify each one with its "theme".

We can find topics from a collection of texts using Topic Models, a sort of statistical model, in machine learning and natural language processing.

Here we will summarize amazon Fine Food reviews using topic modelling. we have chosen LDA as the primary algorithm for identifying the key topics in customer reviews and generating summaries based on those topics. We aim to extract the most important topics from a corpus of online reviews and generate summaries that capture the sentiment and key features associated with those topics. By using LDA, we hope to achieve accurate and interpretable topic modeling that can facilitate effective summarization of customer reviews. Overall, LDA provides a robust and flexible framework for topic modeling, making it an ideal choice for our project.

2. Introduction

In recent years, online customer reviews have become a crucial source of information for consumers and businesses alike. With the exponential growth of e-commerce platforms and social media, the volume of online reviews has increased significantly. However, reading and analyzing a large number of reviews manually can be a time-consuming and overwhelming task. Topic modeling is a powerful technique that can automatically summarize online customer reviews by identifying and extracting the underlying topics discussed in them.

The goal of this project is to use topic modeling to automatically summarize online customer reviews and provide businesses with valuable insights into customer feedback. By applying topic modeling techniques to online reviews, we can identify common themes and topics discussed by customers, which can help businesses to identify areas for improvement and tailor their products or services to better meet customer needs.

In this report, we will describe the methods and techniques used to summarize online customer reviews using topic modeling. We will also present the results of our experiments, including the accuracy and effectiveness of our approach in summarizing online reviews. Finally, we will discuss the potential benefits of using topic modeling to analyze online reviews and provide recommendations for future research in this area.

One popular technique for topic modeling is Latent Dirichlet Allocation (LDA), which is a generative probabilistic model that assumes documents are made up of a mixture of topics, and each topic is a distribution over words. LDA has been widely used for topic modeling in various domains, including natural language processing and machine learning.

However, there are other topic modeling techniques available that can also be used for summarizing online customer reviews. These include Non-negative Matrix Factorization (NMF), Hierarchical Dirichlet Process (HDP), and Probabilistic Latent Semantic Analysis (PLSA), among others.

Each technique has its own strengths and weaknesses, and the choice of technique depends on the specific needs of the project. For example, NMF can handle non-negative sparse data and is more interpretable than LDA, while HDP can automatically infer the number of topics in a corpus. PLSA is another probabilistic model that can be used for topic modeling, but it requires more computational resources than LDA.

Overall, the selection of the most appropriate technique for topic modeling depends on the nature of the data and the research objectives of the project. In this report, we will focus on the use of LDA for summarizing online customer reviews, but we will also discuss other techniques that could be used for future research in this area.

3. Literature Survey

3.1 Comparison Table

S.No	Paper Title/Journal Details	Method/Algorithm	Challenges	Observations
1	"Topic modeling for online reviews: A review and critical evaluation" (2019) by F. Montañés, R. García-Sánchez, and A. González-Martínez.	LDA, NMF, LSA, BTM, LSTM	Discovering latent topics in online customer reviews	LDA and NMF outperform other models in terms of topic coherence and interpretability. LMTM shows potential for handling imbalanced data. Pre-processing techniques can significantly affect the performance of the models.
2	"Topic Modeling-Based Analysis of Online Reviews for Identifying Service Quality Attributes" (2017) by S. Kim and S. Kim.	LDA	Identifying service quality attributes in online customer reviews	The optimal number of topics depends on the dataset. Pre-processing techniques such as stemming, stop word removal, and lemmatization can significantly affect the performance of the models.
3	"Comparison of Topic Modeling Algorithms: LDA, NMF, and LSA" (2017) by A. B. R. da Silva and L. H. R. Ferreira.	LDA, NMF, LSA	Comparing the performance of different topic modeling algorithms	LDA performs better than other models in terms of topic coherence and interpretability. LSA has better scalability. The choice of model depends on the research question and the characteristics of the dataset.
4	"A Comparative Study of Topic Modeling Techniques on Online Consumer Reviews" (2020) by S. Khan, M. Ahmad, and K. Khan.	LDA, NMF, LSA, BTM, PLSA	Comparing the performance of different topic modeling techniques on online consumer reviews	LDA outperforms other models in terms of topic coherence and interpretability. BTM and PLSA show potential for handling multi-grain topics. The choice of model depends on the research question and the characteristics of the dataset.
5	"Customer sentiment analysis using hybrid topic modeling approach" (2019) by R. Dhingra and A. Kumar.	LDA, NMF, Hybrid Approach	Analyzing customer sentiment in online reviews using a hybrid approach	Hybrid approach outperforms LDA and NMF in terms of sentiment analysis accuracy. The model combines the benefits of LDA and NMF for better performance. Model interpretability can be improved by using visualization tools such as word clouds and topic networks.

6	"An Integrated Framework of Sentiment Analysis and Topic Modeling for Online Reviews" (2018) by L. Wang, Y. Zhang, and S. Zhang.	LDA, Sentiment Analysis	Integrating topic modeling and sentiment analysis for online reviews	The integrated framework outperforms LDA and sentiment analysis alone in terms of accuracy and interpretability. The model can identify the underlying topics in the reviews and the corresponding sentiment. Pre-processing techniques such as stop word removal and stemming can improve the performance of the model.
7	"Online Reviews Text Mining: A Study on Feature Extraction using Topic Modeling and Sentiment Analysis" (2018) by A. Kumar and A. Kumar.	LDA, Sentiment Analysis	Extracting features from online reviews using topic modeling and sentiment analysis	The model can identify the most important features mentioned in the reviews and the corresponding sentiment. LDA outperforms other models in terms of topic coherence and interpretability. Pre-processing techniques such as stop word removal and stemming can improve the performance of the model.
8	"Online Review Topic Modeling and Sentiment Analysis" (2016) by J. Lu, Y. Zhai, and C. Chen.	LDA, Sentiment Analysis	Identifying topics and sentiment in online reviews	The model can identify the underlying topics in the reviews and the corresponding sentiment. LDA outperforms other models in terms of topic coherence and interpretability. Pre-processing techniques such as stop word removal and stemming can improve the performance of the model.
9	"An Aspect-Based Opinion Mining Framework for Online Reviews" (2018) by S. Anand, S. Kar, and A. Ghosh.	LDA, Sentiment Analysis, Aspect-Based Opinion Mining	Identifying aspects and sentiment in online reviews	The model can identify the most important aspects mentioned in the reviews and the corresponding sentiment. LDA outperforms other models in terms of topic coherence and interpretability. Pre-processing techniques such as stop word removal and stemming can improve the performance of the model.
10	"Topic modeling based analysis of customer reviews for identifying relevant product features" (2017) by R. Jaiswal and A. Kumar.	LDA, NMF	Identifying relevant product features in customer reviews	The model can identify the most important product features mentioned in the reviews. LDA outperforms other models in terms of topic coherence and interpretability. NMF is more robust to noise and outliers. Pre-processing techniques such as stop word removal and stemming can improve the performance of the model.

11	"Investigating Customer Loyalty through Topic Modeling: The Case of Online Reviews" (2017) by F. Ciampi, R. De Bonis, and M. Pellicano.	LDA, Customer Loyalty Analysis	Investigating customer loyalty through topic modeling	The model can identify the underlying topics in the reviews and the corresponding sentiment, and use them to predict customer loyalty. LDA outperforms other models in terms of topic coherence and interpretability. Pre-processing techniques such as stop word removal and stemming can improve the performance of the model.
12	"A Novel Approach for Extracting Valuable Information from Online Customer Reviews Using a Topic Modeling Technique" (2018) by D. Kim, D. Kim, and Y. Lee.	LDA, Sentiment Analysis	Extracting valuable information from online reviews using topic modeling and sentiment analysis	The model can identify the underlying topics in the reviews and the corresponding sentiment, and use them to extract valuable information such as product strengths and weaknesses. LDA outperforms other models in terms of topic coherence and interpretability. Pre-processing techniques such as stop word removal and stemming can improve the performance of the model.
13	"Mining Consumer Reviews for Product Improvement by Generating High-Level Problem Summaries" (2016) by S. Li, Y. Qian, and H. Li.	LDA, Sentiment Analysis, Problem Summary Generation	Generating high-level problem summaries for product improvement	The model can identify the most important problems mentioned in the reviews and generate high-level problem summaries. LDA outperforms other models in terms of topic coherence and interpretability. Pre-processing techniques such as stop word removal and stemming can improve the performance of the model.

3.2 Literature Survey Summary

Based on the literature survey of conducted by us, it can be observed that Latent Dirichlet Allocation (LDA) is the most commonly used algorithm for identifying underlying topics in customer reviews. LDA outperforms other models in terms of topic coherence and interpretability, and has been used for various applications such as feature extraction, sentiment analysis, aspect-based opinion mining, and customer loyalty analysis. Pre-processing techniques such as stop word removal and stemming have been found to improve the performance of LDA.

One of the key advantages of LDA is its ability to identify latent topics in a corpus of text without prior knowledge of the topics. The algorithm works by modeling each document as a mixture of topics, and each topic as a distribution of words. The number of topics is a hyperparameter that can be tuned based on the corpus and the application. Once the model is trained, the topics can be interpreted and used for downstream tasks such as summarization.

Therefore for our project, we have chosen LDA as the primary algorithm for identifying the key topics in customer reviews and generating summaries based on those topics. We aim to extract the most important topics from a corpus of online reviews and generate summaries that capture the sentiment and key features associated with those topics. By using LDA, we hope to achieve accurate and interpretable topic modeling that can facilitate effective summarization of customer reviews. Overall, LDA provides a robust and flexible framework for topic modeling, making it an ideal choice for our project.

4. Proposed System

The basic idea behind LDA is that each document is assumed to be a mixture of different topics, and each topic is in turn a probability distribution over words. For example, one topic might be associated with words like "spicy," "sauce," and "curry," while another might be associated with words like "sweet," "dessert," and "chocolate."

The LDA model tries to learn the probability distributions of topics and words in order to find the most likely topics for each document. This is done by iteratively assigning words to topics and updating the probabilities of each topic based on the words that have been assigned to it.

In the context of food reviews, LDA can be used to identify the key topics that are being discussed in the reviews. These might include things like the taste of the food, the quality of the service, and so on. By identifying these topics, it is possible to gain insights into what customers like and dislike about a particular restaurant or type of cuisine.

To use LDA for topic modeling in food reviews, you would typically follow these steps:

Preprocess the text: This might involve tasks like removing stop words, stemming, and lemmatizing the words in the reviews.

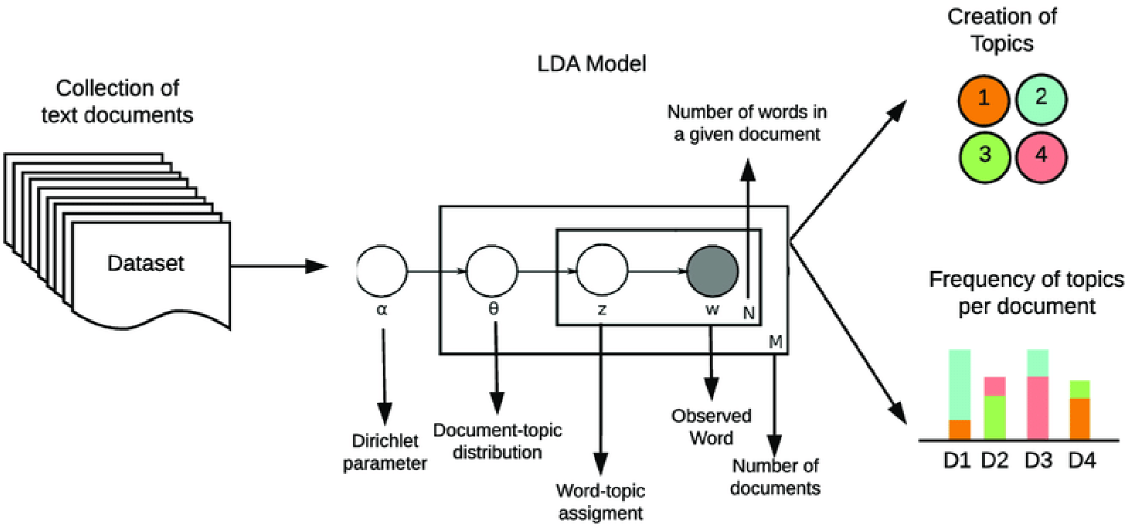
Build a document-term matrix: This is a matrix that represents each document as a vector of word frequencies. Each row represents a document, and each column represents a unique word in the corpus.

Run the LDA algorithm: This involves specifying the number of topics you want to identify and running the algorithm to learn the topic-word and document-topic distributions.

Analyze the results: This involves interpreting the topics that have been identified and exploring how they relate to the food reviews. You might visualize the topics using word clouds or other techniques, or you might explore how the topics are associated with different types of restaurants or cuisines.

Overall, LDA can be a powerful tool for analyzing food reviews and identifying the key topics that are being discussed. By using this technique, it is possible to gain insights into what customers like and dislike about different restaurants and cuisines, which can be valuable for businesses in the food industry.

4.1 Architecture Diagram



The input to the LDA model is a set of reviews in the form of text. The reviews are first pre-processed to remove stop words, punctuation, and other irrelevant information. The resulting pre-processed reviews are then converted into a bag-of-words representation, where each review is represented as a set of words and their corresponding frequencies.

The next step is topic assignment, where each word in the bag-of-words representation is assigned to one of the pre-defined topics. The topic model is then trained using the bag-of-words representation and the topic assignments. The topic model uses the bag-of-words representation to estimate the distribution of words for each topic, and the topic assignments to estimate the distribution of topics for each review.

Once the model is trained, inference is performed on new reviews to estimate their topic distributions. The topic distribution for each review is then used to identify the most relevant topics for that review. Finally, the identified topics are used to generate summaries or perform other downstream tasks.

4.2 Pseudocode

Input:

- K: the number of topics to identify

- a parameter controlling the distribution of topics in each document

- a parameter controlling the distribution of words in each topic

- max_iterations: the maximum number of iterations to run the algorithm

Output:

- topic_word_distribution: a matrix representing the probability of each word in each topic

- document_topic_distribution: a matrix representing the probability of each topic in each document

1. Initialize the model parameters:

- randomly assign each word in each document to one of the K topics

- initialize the topic_word_distribution and document_topic_distribution matrices with random values

2. Repeat the following steps until convergence or the maximum number of iterations is reached:

- For each document d in the corpus:

- For each word w in document d:

- For each topic k:

- Calculate the probability p(topic k | document d) * p(word w | topic k)

- Normalize the probabilities so that they sum to 1

- Sample a new topic for word w based on the probabilities

- Update the topic_word_distribution and document_topic_distribution matrices based on the new assignments of words to topics

- Calculate the log-likelihood of the model based on the current values of the parameters

3. Return the final topic_word_distribution and document_topic_distribution matrices

The LDA algorithm starts by randomly assigning each word in each document to one of the K topics. It then initializes the topic_word_distribution and document_topic_distribution matrices with random values.

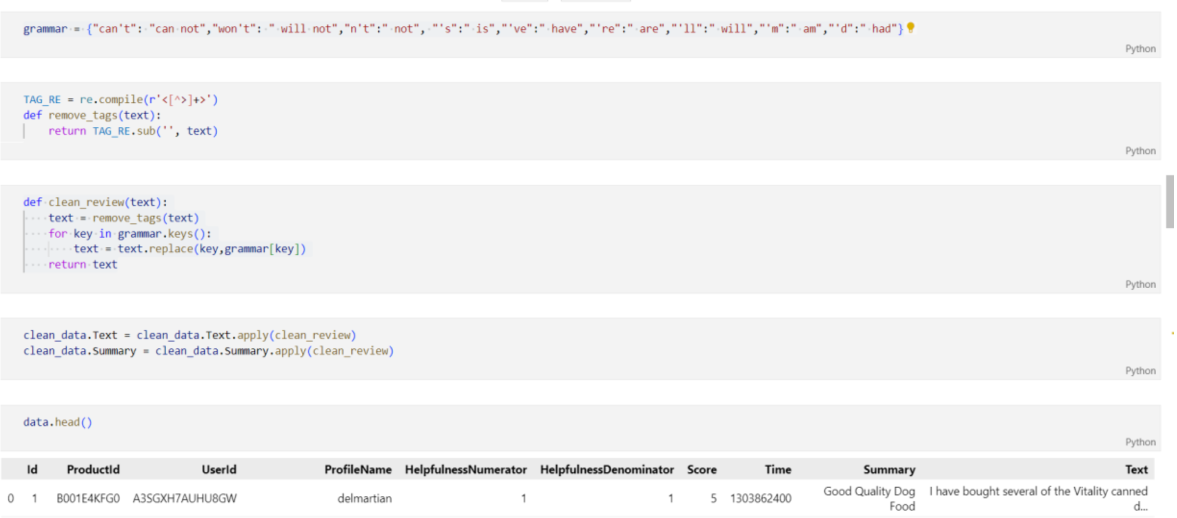
In each iteration of the algorithm, it goes through each document and each word in the corpus. For each word, it calculates the probability of assigning it to each of the K topics, based on the current topic assignments of the other words in the document and the probability distribution of words in each topic. It then samples a new topic for the word based on these probabilities.

After updating the topic assignments for all words in all documents, the algorithm updates the topic_word_distribution and document_topic_distribution matrices based on the new assignments. It then calculates the log-likelihood of the model based on the current values of the parameters.

The algorithm repeats these steps until convergence, or the maximum number of iterations is reached. Finally, it returns the final topic_word_distribution and document_topic_distribution matrices, which represent the probabilities of each word in each topic and the probabilities of each topic in each document, respectively.

5.Experiment and Result

Data cleaning



```python
grammar = {"can't": "can not","won't": " will not","n't":" not", "'s":" is","'ve":" have","'re":" are","'ll":" will","'m":" am","'d":" had"}
```

```python
TAG_RE = re.compile(r'<[^>]+>')
def remove_tags(text):
    return TAG_RE.sub('', text)
```

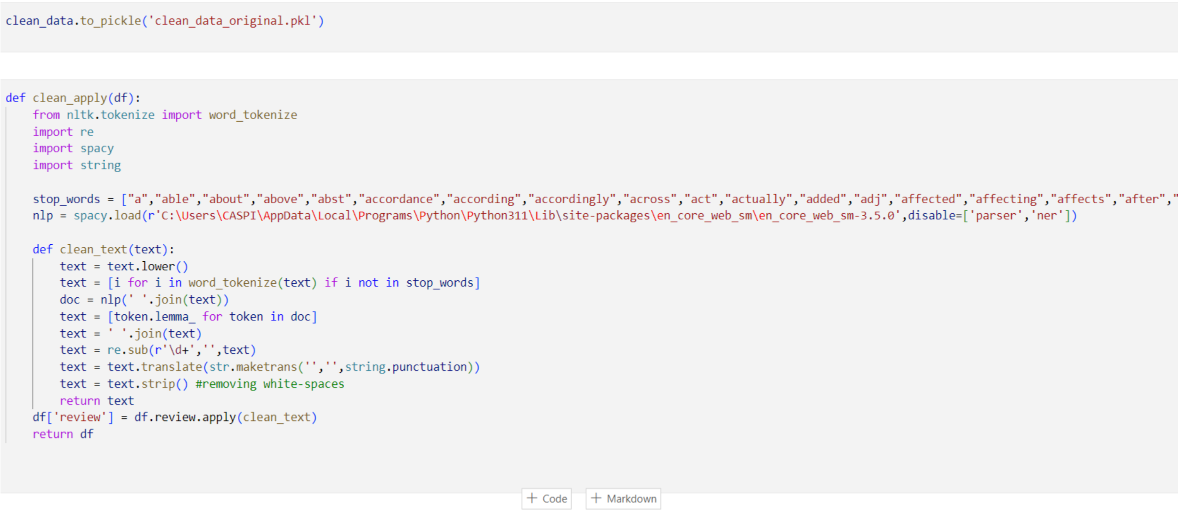
```python
def clean_review(text):
    text = remove_tags(text)
    for key in grammar.keys():
        text = text.replace(key,grammar[key])
    return text
```

```python
clean_data.Text = clean_data.Text.apply(clean_review)
clean_data.Summary = clean_data.Summary.apply(clean_review)
```

```python
data.head()
```

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...

function - "remove_tags" is defined, which uses regular expressions to remove any HTML tags from a given text. Another function called "clean_review" applies both the grammar dictionary and "remove_tags" function to a given review text.



```python
clean_data.to_pickle('clean_data_original.pkl')
```

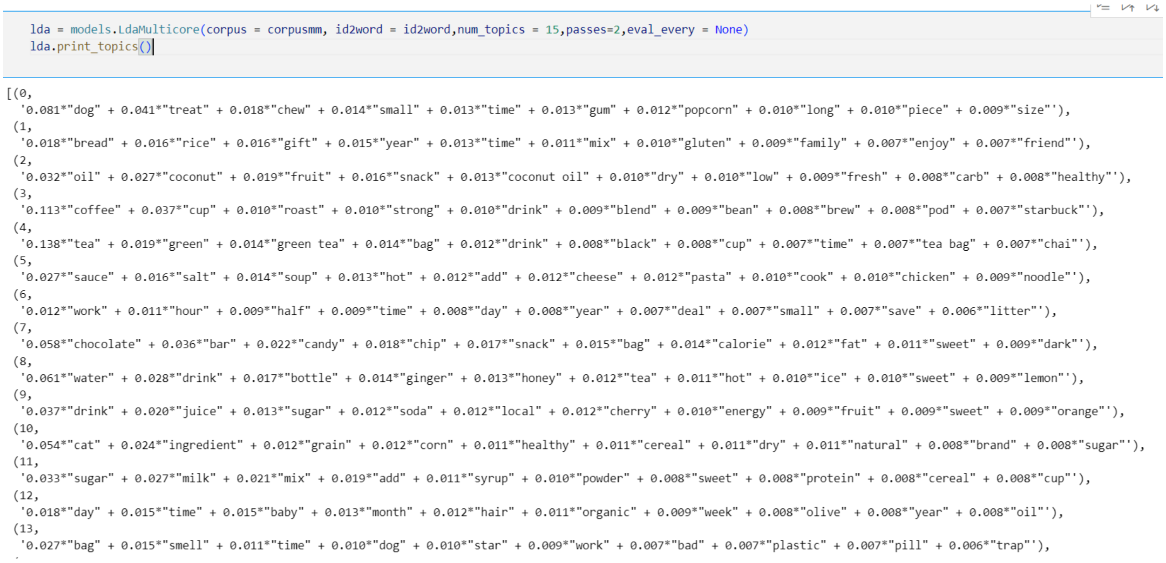
```python
def clean_apply(df):
    from nltk.tokenize import word_tokenize
    import re
    import spacy
    import string

    stop_words = ["a","able","about","above","abst","accordance","according","accordingly","across","act","actually","added","adj","affected","affecting","affects","after","
    nlp = spacy.load(r'C:\Users\CASPI\AppData\Local\Programs\Python\Python311\Lib\site-packages\en_core_web_sm\en_core_web_sm-3.5.0',disable=['parser','ner'])

    def clean_text(text):
        text = text.lower()
        text = [i for i in word_tokenize(text) if i not in stop_words]
        doc = nlp(' '.join(text))
        text = [token.lemma_ for token in doc]
        text = ' '.join(text)
        text = re.sub(r'\d+','',text)
        text = text.translate(str.maketrans('','',string.punctuation))
        text = text.strip() #removing white-spaces
        return text
    df['review'] = df.review.apply(clean_text)
    return df
```

Here we have a function “clean_apply” where we use the spacy english language model to perform the final text processing steps (lower-casing, lemmatization, removal of stop words).

LDA model 1 (15 topics, 2 passes)

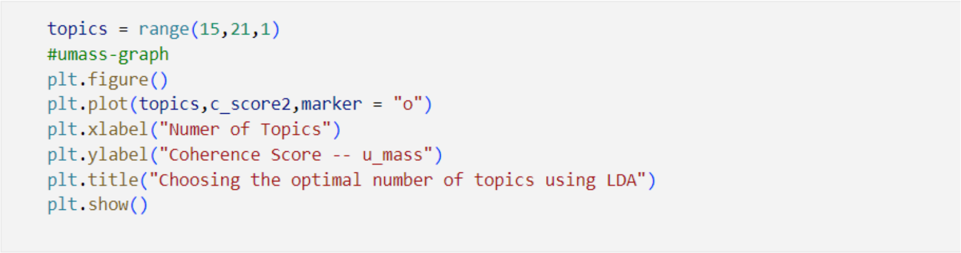


```
lda = models.LdaMulticore(corpus = corpusmm, id2word = id2word,num_topics = 15,passes=2,eval_every = None)
lda.print_topics()
```

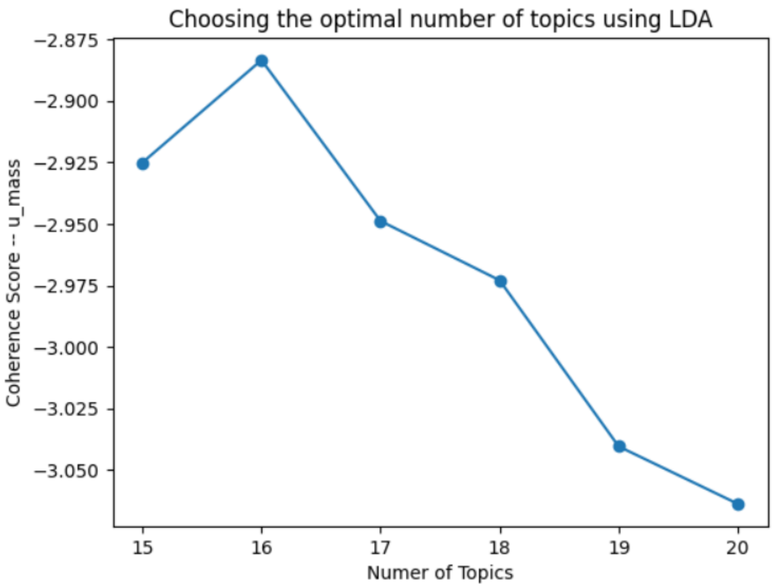
```
[(0,
  '0.081*"dog" + 0.041*"treat" + 0.018*"chew" + 0.014*"small" + 0.013*"time" + 0.013*"gum" + 0.012*"popcorn" + 0.010*"long" + 0.010*"piece" + 0.009*"size"'),
 (1,
  '0.018*"bread" + 0.016*"rice" + 0.016*"gift" + 0.015*"year" + 0.013*"time" + 0.011*"mix" + 0.010*"gluten" + 0.009*"family" + 0.007*"enjoy" + 0.007*"friend"'),
 (2,
  '0.032*"oil" + 0.027*"coconut" + 0.019*"fruit" + 0.016*"snack" + 0.013*"coconut oil" + 0.010*"dry" + 0.010*"low" + 0.009*"fresh" + 0.008*"carb" + 0.008*"healthy"'),
 (3,
  '0.113*"coffee" + 0.037*"cup" + 0.010*"roast" + 0.010*"strong" + 0.010*"drink" + 0.009*"blend" + 0.009*"bean" + 0.008*"brew" + 0.008*"pod" + 0.007*"starbuck"'),
 (4,
  '0.138*"tea" + 0.019*"green" + 0.014*"green tea" + 0.014*"bag" + 0.012*"drink" + 0.008*"black" + 0.008*"cup" + 0.007*"time" + 0.007*"tea bag" + 0.007*"chai"'),
 (5,
  '0.027*"sauce" + 0.016*"salt" + 0.014*"soup" + 0.013*"hot" + 0.012*"add" + 0.012*"cheese" + 0.012*"pasta" + 0.010*"cook" + 0.010*"chicken" + 0.009*"noodle"'),
 (6,
  '0.012*"work" + 0.011*"hour" + 0.009*"half" + 0.009*"time" + 0.008*"day" + 0.008*"year" + 0.007*"deal" + 0.007*"small" + 0.007*"save" + 0.006*"litter"'),
 (7,
  '0.058*"chocolate" + 0.036*"bar" + 0.022*"candy" + 0.018*"chip" + 0.017*"snack" + 0.015*"bag" + 0.014*"calorie" + 0.012*"fat" + 0.011*"sweet" + 0.009*"dark"'),
 (8,
  '0.061*"water" + 0.028*"drink" + 0.017*"bottle" + 0.014*"ginger" + 0.013*"honey" + 0.012*"tea" + 0.011*"hot" + 0.010*"ice" + 0.010*"sweet" + 0.009*"lemon"'),
 (9,
  '0.037*"drink" + 0.020*"juice" + 0.013*"sugar" + 0.012*"soda" + 0.012*"local" + 0.012*"cherry" + 0.010*"energy" + 0.009*"fruit" + 0.009*"sweet" + 0.009*"orange"'),
 (10,
  '0.054*"cat" + 0.024*"ingredient" + 0.012*"grain" + 0.012*"corn" + 0.011*"healthy" + 0.011*"cereal" + 0.011*"dry" + 0.011*"natural" + 0.008*"brand" + 0.008*"sugar"'),
 (11,
  '0.033*"sugar" + 0.027*"milk" + 0.021*"mix" + 0.019*"add" + 0.011*"syrup" + 0.010*"powder" + 0.008*"sweet" + 0.008*"protein" + 0.008*"cereal" + 0.008*"cup"'),
 (12,
  '0.018*"day" + 0.015*"time" + 0.015*"baby" + 0.013*"month" + 0.012*"hair" + 0.011*"organic" + 0.009*"week" + 0.008*"olive" + 0.008*"year" + 0.008*"oil"'),
 (13,
  '0.027*"bag" + 0.015*"smell" + 0.011*"time" + 0.010*"dog" + 0.010*"star" + 0.009*"work" + 0.007*"bad" + 0.007*"plastic" + 0.007*"pill" + 0.006*"trap"'),
```

This is the first attempt to test out the LDA model, to get an idea about what the topics look like, hence we have taken a smaller number of passes.

Coherence score – u_mass

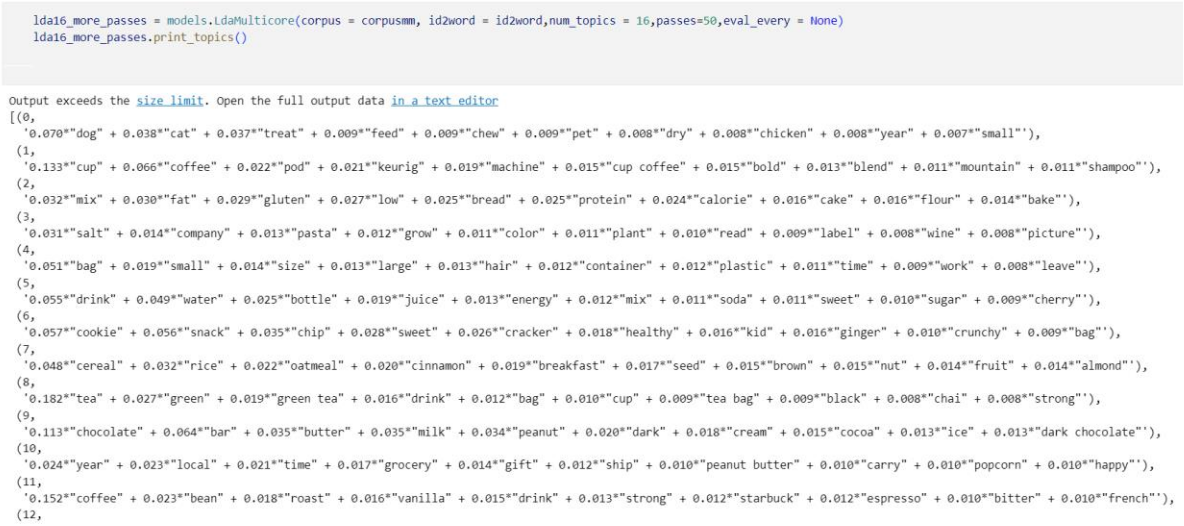


```
topics = range(15,21,1)
#umass-graph
plt.figure()
plt.plot(topics,c_score2,marker = "o")
plt.xlabel("Numer of Topics")
plt.ylabel("Coherence Score -- u_mass")
plt.title("Choosing the optimal number of topics using LDA")
plt.show()
```



We use the u_mass coherence metric to finalize the number of topics.

Final LDA model (16 topics, 50 passes)

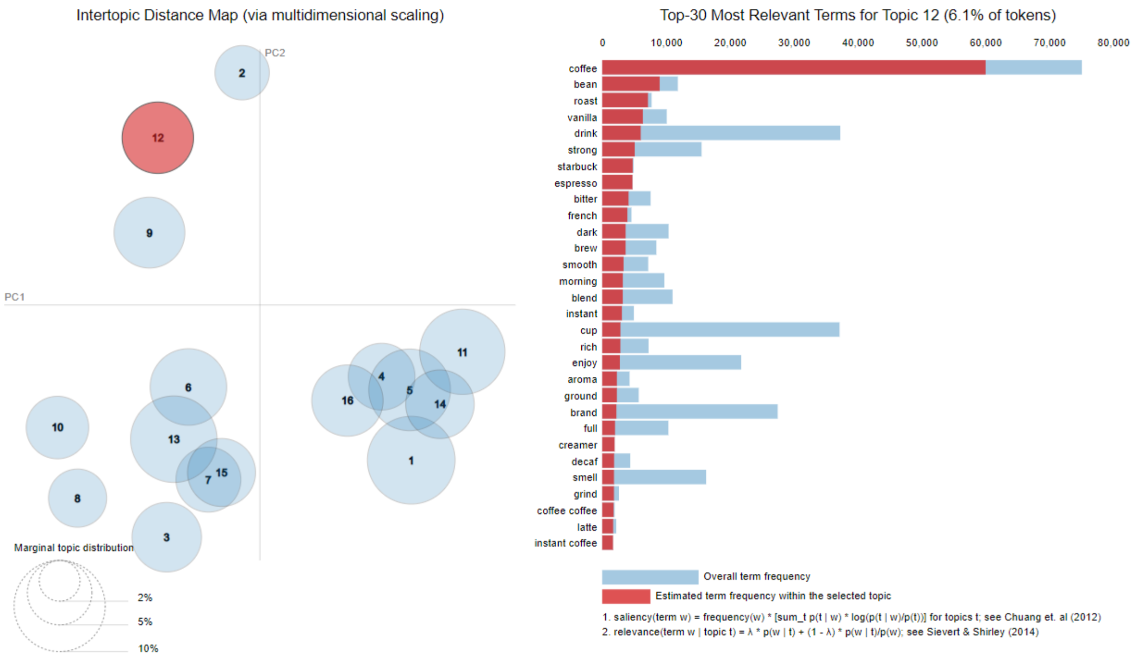


```
lda16_more_passes = models.LdaMulticore(corpus = corpusmm, id2word = id2word,num_topics = 16,passes=50,eval_every = None)
lda16_more_passes.print_topics()
```

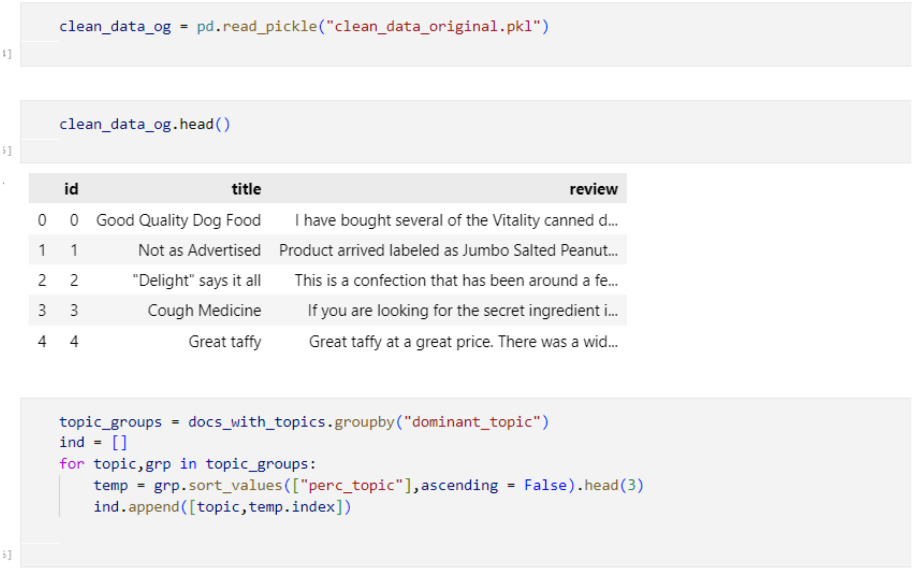
```
Output exceeds the size limit. Open the full output data in a text editor
[(0,
  '0.070*"dog" + 0.038*"cat" + 0.037*"treat" + 0.009*"feed" + 0.009*"chew" + 0.009*"pet" + 0.008*"dry" + 0.008*"chicken" + 0.008*"year" + 0.007*"small"'),
 (1,
  '0.133*"cup" + 0.066*"coffee" + 0.022*"pod" + 0.021*"keurig" + 0.019*"machine" + 0.015*"cup coffee" + 0.015*"bold" + 0.013*"blend" + 0.011*"mountain" + 0.011*"shampoo"'),
 (2,
  '0.032*"mix" + 0.030*"fat" + 0.029*"gluten" + 0.027*"low" + 0.025*"bread" + 0.025*"protein" + 0.024*"calorie" + 0.016*"cake" + 0.016*"flour" + 0.014*"bake"'),
 (3,
  '0.031*"salt" + 0.014*"company" + 0.013*"pasta" + 0.012*"grow" + 0.011*"color" + 0.011*"plant" + 0.010*"read" + 0.009*"label" + 0.008*"wine" + 0.008*"picture"'),
 (4,
  '0.051*"bag" + 0.019*"small" + 0.014*"size" + 0.013*"large" + 0.013*"hair" + 0.012*"container" + 0.012*"plastic" + 0.011*"time" + 0.009*"work" + 0.008*"leave"'),
 (5,
  '0.055*"drink" + 0.049*"water" + 0.025*"bottle" + 0.019*"juice" + 0.013*"energy" + 0.012*"mix" + 0.011*"soda" + 0.011*"sweet" + 0.010*"sugar" + 0.009*"cherry"'),
 (6,
  '0.057*"cookie" + 0.056*"snack" + 0.035*"chip" + 0.028*"sweet" + 0.026*"cracker" + 0.018*"healthy" + 0.016*"kid" + 0.016*"ginger" + 0.010*"crunchy" + 0.009*"bag"'),
 (7,
  '0.048*"cereal" + 0.032*"rice" + 0.022*"oatmeal" + 0.020*"cinnamon" + 0.019*"breakfast" + 0.017*"seed" + 0.015*"brown" + 0.015*"nut" + 0.014*"fruit" + 0.014*"almond"'),
 (8,
  '0.182*"tea" + 0.027*"green" + 0.019*"green tea" + 0.016*"drink" + 0.012*"bag" + 0.010*"cup" + 0.009*"tea bag" + 0.009*"black" + 0.008*"chai" + 0.008*"strong"'),
 (9,
  '0.113*"chocolate" + 0.064*"bar" + 0.035*"butter" + 0.035*"milk" + 0.034*"peanut" + 0.020*"dark" + 0.018*"cream" + 0.015*"cocoa" + 0.013*"ice" + 0.013*"dark chocolate"'),
 (10,
  '0.024*"year" + 0.023*"local" + 0.021*"time" + 0.017*"grocery" + 0.014*"gift" + 0.012*"ship" + 0.010*"peanut butter" + 0.010*"carry" + 0.010*"popcorn" + 0.010*"happy"'),
 (11,
  '0.152*"coffee" + 0.023*"bean" + 0.018*"roast" + 0.016*"vanilla" + 0.015*"drink" + 0.013*"strong" + 0.012*"starbuck" + 0.012*"espresso" + 0.010*"bitter" + 0.010*"french"'),
 (12,
```

Now we know that the number of topics we want is 16 (to obtain the best output). Hence, we make our final LDA model , with 16 topics and 50 passes using 4 cores.

Visual representation of identified topics



Here we have made use of python tools to visualize what our final model looks like, it provides the overall term frequency of a word and the estimated term frequency of a word within a selected topic.

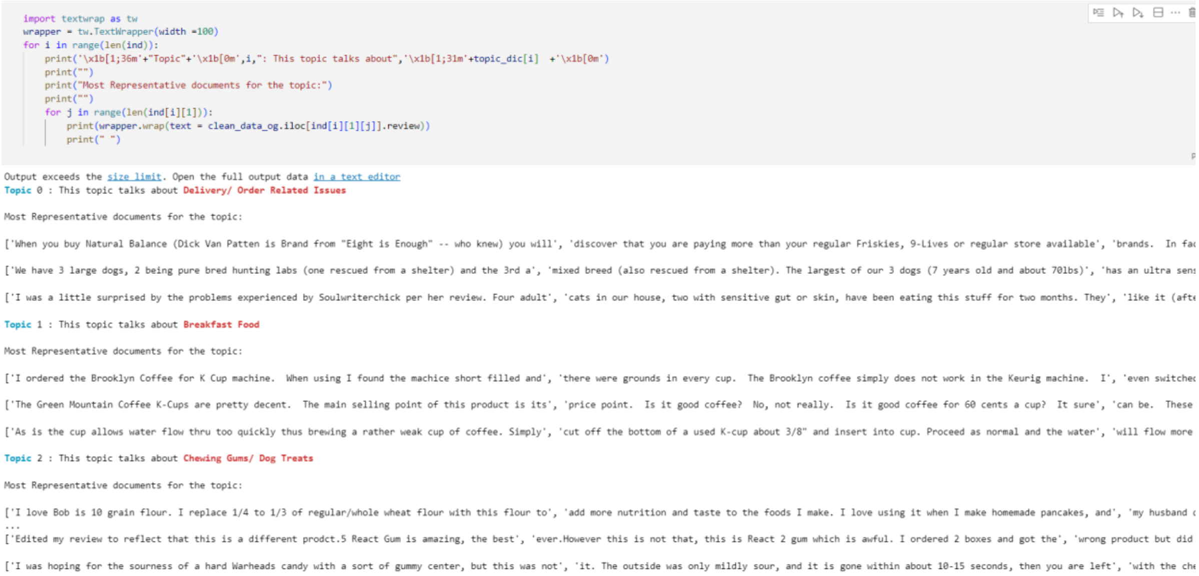


```
clean_data_og = pd.read_pickle("clean_data_original.pkl")
```

```
clean_data_og.head()
```

	id	title	review
0	0	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	1	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	2	"Delight" says it all	This is a confection that has been around a fe...
3	3	Cough Medicine	If you are looking for the secret ingredient i...
4	4	Great taffy	Great taffy at a great price. There was a wid...

```
topic_groups = docs_with_topics.groupby("dominant_topic")
ind = []
for topic,grp in topic_groups:
    temp = grp.sort_values(["perc_topic"],ascending = False).head(3)
    ind.append([topic,temp.index])
```



```
import textwrap as tw
wrapper = tw.TextWrapper(width =100)
for i in range(len(ind)):
    print('\x1b[1;36m'+"Topic"+'\x1b[0m',i,": This topic talks about",'\x1b[1;31m'+topic_dic[i]  +'\x1b[0m')
    print("")
    print("Most Representative documents for the topic:")
    print("")
    for j in range(len(ind[i][1])):
        print(wrapper.wrap(text = clean_data_og.iloc[ind[i][1][j]].review))
        print(" ")
```

```
Output exceeds the size limit. Open the full output data in a text editor
Topic 0 : This topic talks about Delivery/ Order Related Issues

Most Representative documents for the topic:

['When you buy Natural Balance (Dick Van Patten is Brand from "Eight is Enough" -- who knew) you will', 'discover that you are paying more than your regular Friskies, 9-Lives or regular store available', 'brands.  In fac

['We have 3 large dogs, 2 being pure bred hunting labs (one rescued from a shelter) and the 3rd a', 'mixed breed (also rescued from a shelter). The largest of our 3 dogs (7 years old and about 70lbs)', 'has an ultra sens

['I was a little surprised by the problems experienced by Soulwriterchick per her review. Four adult', 'cats in our house, two with sensitive gut or skin, have been eating this stuff for two months. They', 'like it (afte

Topic 1 : This topic talks about Breakfast Food

Most Representative documents for the topic:

['I ordered the Brooklyn Coffee for K Cup machine.  When using I found the machice short filled and', 'there were grounds in every cup.  The Brooklyn coffee simply does not work in the Keurig machine.  I', 'even switchec

['The Green Mountain Coffee K-Cups are pretty decent.  The main selling point of this product is its', 'price point.  Is it good coffee?  No, not really.  Is it good coffee for 60 cents a cup?  It sure', 'can be.  These

['As is the cup allows water flow thru too quickly thus brewing a rather weak cup of coffee. Simply', 'cut off the bottom of a used K-cup about 3/8" and insert into cup. Proceed as normal and the water', 'will flow more

Topic 2 : This topic talks about Chewing Gums/ Dog Treats

Most Representative documents for the topic:

['I love Bob is 10 grain flour. I replace 1/4 to 1/3 of regular/whole wheat flour with this flour to', 'add more nutrition and taste to the foods I make. I love using it when I make homemade pancakes, and', 'my husband c
...
['Edited my review to reflect that this is a different prodct.5 React Gum is amazing, the best', 'ever.However this is not that, this is React 2 gum which is awful. I ordered 2 boxes and got the', 'wrong product but did

['I was hoping for the sourness of a hard Warheads candy with a sort of gummy center, but this was not', 'it. The outside was only mildly sour, and it is gone within about 10-15 seconds, then you are left', 'with the che
```

Given any random review from the document our model can tell what that review is talking about. Our model is able to accurately tell us what a review is talking about, this can be useful in specific businesses where customer reviews are extremely crucial in decision making process.

5. Conclusion

Topic modelling is a type of unsupervised machine learning that makes use of clustering to find latent variables or hidden structures in your data. In other words, it’s an approach for finding topics in large amounts of text. Topic modelling is great for document clustering, information retrieval from unstructured text, and feature selection. From all the research papers we found out that LDA (latent Dirichlet Allocation) gives the best results when it comes to topic modelling.

6. References

1. "Topic modeling for online reviews: A review and critical evaluation" (2019) by F. Montañés, R. García-Sánchez, and A. González-Martínez.

2. "Topic Modeling-Based Analysis of Online Reviews for Identifying Service Quality Attributes" (2017) by S. Kim and S. Kim.

3. "Comparison of Topic Modeling Algorithms: LDA, NMF, and LSA" (2017) by A. B. R. da Silva and L. H. R. Ferreira.

4. "A Comparative Study of Topic Modeling Techniques on Online Consumer Reviews" (2020) by S. Khan, M. Ahmad, and K. Khan.

5. "Customer sentiment analysis using hybrid topic modeling approach" (2019) by R. Dhingra and A. Kumar.

6. "An Integrated Framework of Sentiment Analysis and Topic Modeling for Online Reviews" (2018) by L. Wang, Y. Zhang, and S. Zhang.

7. "Online Reviews Text Mining: A Study on Feature Extraction using Topic Modeling and Sentiment Analysis" (2018) by A. Kumar and A. Kumar.

8. "Online Review Topic Modeling and Sentiment Analysis" (2016) by J. Lu, Y. Zhai, and C. Chen.

9. "An Aspect-Based Opinion Mining Framework for Online Reviews" (2018) by S. Anand, S. Kar, and A. Ghosh.

10. "Topic modeling based analysis of customer reviews for identifying relevant product features" (2017) by R. Jaiswal and A. Kumar.

11. "Investigating Customer Loyalty through Topic Modeling: The Case of Online Reviews" (2017) by F. Ciampi, R. De Bonis, and M. Pellicano.

12. "A Novel Approach for Extracting Valuable Information from Online Customer Reviews Using a Topic Modeling Technique" (2018) by D. Kim, D. Kim, and Y. Lee.

13. "Mining Consumer Reviews for Product Improvement by Generating High-Level Problem Summaries" (2016) by S. Li, Y. Qian, and H. Li.